Source Localization in Reverberant Environments: Modeling and Statistical Analysis

Tony Gustafsson, *Member, IEEE*, Bhaskar D. Rao, *Fellow, IEEE*, and Mohan Trivedi, *Senior Member, IEEE*

***Abstract*—Room reverberation is typically the main obstacle for designing robust microphone-based source localization systems. The purpose of the paper is to analyze the achievable performance of acoustical source localization methods when room reverberation is present.**

To facilitate the analysis, we apply well known results from room acoustics to develop a simple but useful statistical model for the room transfer function. The properties of the statistical model are found to correlate well with results from real data measurements.

The room transfer function model is further applied to analyze the statistical properties of some existing methods for source localization. In this respect we consider especially the asymptotic error variance and the probability of an anomalous estimate. A noteworthy outcome of the analysis is that the so-called PHAT time-delay estimator is shown to be optimal among a class of cross-correlation based time-delay estimators. To verify our results on the error variance and the outlier probability we apply the image method for simulation of the room transfer function.

***Index Terms*—Acoustic arrays, acoustic signal processing.**

I. Introduction

Several approaches for acoustical source localization using microphone arrays have appeared in the literature. Among existing proposals, those based on time-delay estimation (TDE) have gained the most attention, see e.g., [1]–[4]. In this context, the signals received from a pair of microphones are modeled as

$$\begin{aligned} x_1(t) &= s(t) + n_1(t) \\ x_2(t) &= s(t-\tau_0) + n_2(t) \end{aligned} \tag{1}$$

where $x_i(t)$ $(i = 1, 2)$ is the output of the ith receiver, $s(t)$ is the unknown source signal, $n_i(t)$ is an additive noise term assumed uncorrelated with $s(t)$, and τ_0 is the difference in propagation times (i.e., the unknown relative *time-delay*). Assuming that several such microphone pairs are distributed over the spatial region, the source location is obtained from the estimated time-delays, see, e.g., [2], [4], [5].

Knowledge about the location of the source is crucial e.g., for automatic speaker tracking in video-conferencing [6], [7]. In office-like environments, the accuracy of the estimated time-delay is typically limited by *room reverberation* rather than by additive uncorrelated noise as in (1). In a reverberant environment, the measured microphone signals are modeled as

$$\mathbf{x}(t) = \int_{-\infty}^{\infty} \mathbf{h}(t-\lambda)s(\lambda)\,d\lambda + \mathbf{n}(t) \tag{2}$$

where $\mathbf{x}(t) = [x_1(t)\ \ x_2(t)]^T, \mathbf{h}(t) = [h_1(t)\ \ h_2(t)]^T$ and $\mathbf{n}(t) = [n_1(t)\ \ n_2(t)]^T$. Here, $h_i(t)$ represents the impulse response of the acoustical transfer function from the source to the ith microphone. To facilitate the analysis we assume that $\mathbf{h}(t)$ is time-invariant; it is hence assumed that the source location is fixed.

The empirical experience is that once the level of room reverberation rises above minimal levels, existing methods for TDE begin to exhibit substantial performance degradations. Among the few publications available concerning the performance of TDE in reverberant environments, we mention [8]–[10]. In [9], Ianniello studied the case with one source and two or three resolvable propagation paths. However, room reverberation typically consists of the superposition of a large number of echoes. The results in [9] are therefore applicable only in situations where the early reflections are strong and "late reverberation" weak. In [8], Champagne *et al.* applied the image method for simulation of $\mathbf{h}(t)$. Based on an analogy with performance bounds for the single-path propagation model (1), a lower bound was suggested for the variance of the estimated time-delay. Although the proposed lower bound in a simulation study accurately predicted the variance of the estimated time-delay, a theoretical justification of the proposed lower bound was not provided. Reference [10] also focuses on the variance of the estimated time-delay, and can be considered as an early version of the present paper.

The main contributions of the present paper are as follows.

1) Utilizing results from statistical room acoustics we develop a room transfer function model that can be of utility for signal processing researchers in developing and analyzing microphone array based source localization algorithms.
2) The statistical model is used to develop Cramér-Rao lower bounds (CRB) for time delay estimates in reverberant environments thereby facilitating an understanding of accuracy limits. Furthermore, these results provide a theoretical basis for results recently developed in this context by [8].
3) The acoustical modeling is also utilized to better understand cross-correlation based methods for time-delay estimation. In particular, the Generalized Cross Correlation

Manuscript received October 5, 2000; revised March 12, 2003. This work was supported by the Swedish Foundation for International Cooperation in Research and Higher Education, Telefonaktiebolaget LM Ericsson, and the University of California DiMi Program D97-17. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Brandstein.

T. Gustafsson is with NIRA Dynamics AB, SE-421 31 Vastra Frolunda, Sweden (e-mail: 031.276711@telia.com).

B. D. Rao and M. Trivedi are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407 USA (e-mail: brao@ece.ucsd.edu; trivedi@ece.ucsd.edu).

Digital Object Identifier 10.1109/TSA.2003.818027

(GCC) method [16] is examined and optimal weightings are determined. Interestingly it turns out that use of the optimal weights results in the PHAse Transform (PHAT) method.

4) The derived CRB is rather local in the sense that it is reachable only under ideal conditions. In practical applications the limiting factor is instead the fact that GCC-based localization methods suffer from outliers when reverberation is present. Assuming that a GCC-based localization method is applied, we derive expressions for the outlier probability. These results should be of much use in understanding how to combine microphone arrays and room acoustics for achieving desired performance.

II. Statistical Room Reverberation Model

In this section, we introduce the room reverberation model, which is built on some well-known results from statistical room acoustics. The key assumption we make is that the impulse response $\mathbf{h}(t)$ can be decomposed as

$$\mathbf{h}(t) = \mathbf{h}_d(t) + \mathbf{r}(t) \tag{3}$$

where subscript $(\cdot)_d$ denotes direct-path propagation, and $\mathbf{r}(t)$ corresponds to *diffuse* sound propagation. Recently, Radlović *et al.* [11] applied a similar model to investigate the robustness of acoustical equalization techniques.

A. Direct Path Propagation

Assume that the source is located at $\mathbf{e}_s$, the microphones are located at $\mathbf{e}_{m_1}$ and $\mathbf{e}_{m_2}$, and let $d = \|\mathbf{e}_{m_1} - \mathbf{e}_{m_2}\|$. The direct-path impulse response is modeled as

$$\mathbf{h}_d(t) = \begin{bmatrix} \delta(t) \\ \delta(t-\tau_0) \end{bmatrix} \kappa(r) \tag{4}$$

where $\delta(\cdot)$ denotes Dirac's delta function, and τ_0 denotes the difference in propagation times

$$\tau_0 = \frac{1}{c}\left(\|\mathbf{e}_s - \mathbf{e}_{m_1}\| - \|\mathbf{e}_s - \mathbf{e}_{m_2}\|\right). \tag{5}$$

Here, the speed of sound is denoted as c (generally specified as $c = 344$ m/s at 21 °C), and $\kappa(r) = 1/\sqrt{4\pi r^2}$, where r is the distance from the source to the point in-between the two microphones. Model (4) is valid when both microphones receive an equal amount of power from the direct-path; it is thus assumed that $d \ll (\|\mathbf{e}_s - \mathbf{e}_{m_1}\|, \|\mathbf{e}_s - \mathbf{e}_{m_2}\|)$ and that the source is a point source.

B. Properties of $\mathbf{r}(t)$

The sound pressure at a microphone is built up of the direct-path, plus several waves due to multiple reflections of the original sound. These reflections travel in different directions and encounter the walls at different angles of incidence. Studying an empty rectangular room, $\mathbf{h}(t)$ can be computed by solving a wave equation, see for example [12, Ch. 3]. At higher frequencies, the complexity (in terms of the number of modes) of "deterministic" wave equation based modeling increases to a point where exact analysis is no longer feasible.

To model the high-frequency part of $\mathbf{h}(t)$, we will apply the theory of random (or diffuse) sound fields. A diffuse sound field is present when the following conditions are fulfilled [12], [13]:

A1: The dimensions of the room are large relative to the wavelength of $s(t)$. For the frequencies of interest (in speech processing we are mainly interested in the band 300–3500 Hz), this condition is usually satisfied.

A2: The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth. In a room with volume V (in m^3), and reverberation time[1] T_{60} (in seconds), this condition is fulfilled for frequencies that exceed the "Schroeder large room frequency":

$$f_S = 2000\sqrt{T_{60}/V}. \tag{6}$$

A3: The source and the microphones are located in the interior of the room, at least a half-wavelength away from the walls. The sound field at a wall-mounted microphone can hence not be modeled as diffuse.

The statistical properties of the transfer function $\mathbf{R}(\omega)$ (the Fourier transform of $\mathbf{r}(t)$) is independent of the time-instant of observation. That is, given a fixed source location and fixed microphone locations, $\mathbf{R}(\omega)$ will remain constant (unless the room configuration somehow is altered!). To emphasize this fact, we define the vector $\boldsymbol{\theta} = [\mathbf{e}_s^T \ \mathbf{e}_{m_1}^T \ \mathbf{e}_{m_2}^T]^T$, and denote the impulse response and its frequency response accordingly: $\mathbf{r}(t;\boldsymbol{\theta}) \leftrightarrow \mathbf{R}(\omega;\boldsymbol{\theta})$. For a fixed $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, we hence consider $\mathbf{r}(t;\tilde{\boldsymbol{\theta}})$ to be a *realization* of the random function $\mathbf{r}(t;\boldsymbol{\theta})$.

Write the kth $(k = 1, 2)$ component of $\mathbf{R}(\omega;\boldsymbol{\theta})$ as

$$R_k(\omega;\boldsymbol{\theta}) = R_k^r(\omega;\boldsymbol{\theta}) + jR_k^i(\omega;\boldsymbol{\theta}) \tag{7}$$

where $R_k^r(\omega;\boldsymbol{\theta})$ and $R_k^i(\omega;\boldsymbol{\theta})$ are the real and imaginary parts of $R_k(\omega;\boldsymbol{\theta})$, respectively, and $j = \sqrt{-1}$. We next cite a couple of useful results from [13] (assuming *A1–A3* to be fulfilled)

$$\varphi_{kk}^{rr}(\Delta\omega) = \varphi_{kk}^{ii}(\Delta\omega) = \frac{1}{1 + \left(\Delta\omega\frac{T_{60}}{13.8}\right)^2}, \tag{8}$$

$$\varphi_{kk}^{ri}(\Delta\omega) = \frac{\Delta\omega\frac{T_{60}}{13.8}}{1 + \left(\Delta\omega\frac{T_{60}}{13.8}\right)^2}. \tag{9}$$

Here, $\varphi_{kk}^{ri}(\Delta\omega)$ denotes the normalized (i.e., $-1 \le \varphi_{kk}^{ri}(\Delta\omega) \le 1$) correlation function between $R_k^r(\omega;\boldsymbol{\theta})$ and $R_k^i(\omega+\Delta\omega;\boldsymbol{\theta})$. *Note, the expectation operator used to define the correlation functions* (8)–(9) *should be interpreted as the ensemble average over all allowable (in terms of assumption* A3*) values of* $\boldsymbol{\theta}$*, denoted with the expectation operator* $E_{\boldsymbol{\theta}}\{\cdot\}$. From [13] we also find that $E_{\boldsymbol{\theta}}\{\mathbf{R}(\omega;\boldsymbol{\theta})\} = \mathbf{0}$.

From (8)–(9), we draw the conclusion that $R_k^r(\omega;\boldsymbol{\theta})$ and $R_k^r(\omega+\Delta\omega;\boldsymbol{\theta})$ are uncorrelated for "$\Delta\omega$ sufficiently large." We next define a "coherence bandwidth" $\rho(T_{60})$ as

$$\rho(T_{60}) \triangleq \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_{kk}^{rr}(\omega)\, d\omega \simeq \frac{7}{T_{60}} \quad [\text{Hz}]. \tag{10}$$

[1] The reverberation time T_{60} is defined as the length of time for the sound intensity level in a room to decrease by 60 dB after the sound source is shut off.

Consequently, when $|\omega_1 - \omega_2|/2\pi \geq 7/T_{60}$ the random variables $R_k^r(\omega_1;\boldsymbol{\theta})$ and $R_k^r(\omega_2;\boldsymbol{\theta})$ are considered uncorrelated.

Consider next the *spatial* correlation between $h_1(t)$ and $h_2(t)$. No general results seem to be available in the literature. However, if $s(t) = \sin(\omega_0 t + \phi)$, and if the conditions for diffuse sound are fulfilled [12, Ch. 8] shows that

$$\frac{E_{\boldsymbol{\theta}}\{x_1(t)x_2(t)\}}{\sqrt{E_{\boldsymbol{\theta}}\left\{x_1^2(t)\right\} E_{\boldsymbol{\theta}}\left\{x_2^2(t)\right\}}} = \operatorname{sinc}\left(\frac{2\pi d}{\lambda_0}\right) \tag{11}$$

where $\lambda_0 = (2\pi c)/\omega_0$, and $\operatorname{sinc}(x) \triangleq \sin(\pi x)/(\pi x)$. If the source energy is not concentrated to a single frequency, ω_0 can be replaced with the mean of the highest and lowest frequencies of the source signal. Expression (11) is only approximately true in that case. Similarly to the definition of the *coherence bandwidth*, we define a *spatial coherence distance* as the first zero-crossing of (11); $\rho_s = \lambda_0/2$. For speech signals, the mean of the highest and lowest frequencies equals approximately $0.5(3500 + 300)$, implying that $\rho_s \simeq 0.1$ m.

Let us finally consider the variance of the random variable $R_k(\omega;\boldsymbol{\theta})$. Reference [12, Ch. 5] shows that

$$\begin{aligned} q^2 &\triangleq E_{\boldsymbol{\theta}}\{R_k(\omega;\boldsymbol{\theta})(R_k(\omega;\boldsymbol{\theta}))^*\} \\ &= \frac{4\beta^2}{\mathcal{A}(1-\beta^2)}. \end{aligned} \tag{12}$$

Here, $(\,\cdot\,)^*$ denotes complex conjugate, β $(0 \leq \beta \leq 1)$ is the reflection coefficient, and $\mathcal{A}$ denotes the total wall area of the room.

C. Model Validation

1) Synthetic Data: We consider an empty rectangular room with plane reflective surfaces. The walls are characterized by the reflection coefficient β, which is assumed identical for all walls. We have simulated a room with dimensions $L_x = 10$, $L_y = 6.6$, and $L_z = 3$ (all measures in meters). We assume that there is an omni-directional (spherical radiator) acoustical point source present in the interior of the room. The radiated signal is measured by omni-directional microphones, which are located in the interior of the room. To generate realizations of $\mathbf{R}(\omega;\boldsymbol{\theta})$, we apply the following procedure.

- Define initial positions of the microphones and the source, and denote them $\tilde{\mathbf{e}}_{m_1}, \tilde{\mathbf{e}}_{m_2}$, and $\tilde{\mathbf{e}}_s$, respectively. The initial positions are chosen such that the desired (r, τ_0, d) are obtained.
- For each simulation, generate a random translation $\mathbf{y}$ and a 3×3 random rotation matrix $\mathbf{G}$ (i.e., $\mathbf{G}^T\mathbf{G} = \mathbf{I}_3$), where $\mathbf{I}_M$ denotes the M-dimensional identity matrix. Let $\mathbf{e}_s = \mathbf{f}(\tilde{\mathbf{e}}_s)$, where $\mathbf{f}(\tilde{\mathbf{e}}_s) = \mathbf{y} + \mathbf{G}\tilde{\mathbf{e}}_s$. The mapping $\mathbf{f}(\,\cdot\,)$ is applied also to $\tilde{\mathbf{e}}_{m_1}$ and $\tilde{\mathbf{e}}_{m_2}$. Note that the triplet (r, τ_0, d) is invariant to $\mathbf{f}(\,\cdot\,)$.
- Apply the image-method [14] to compute $\mathbf{h}(t)$ using $\mathbf{e}_s, \mathbf{e}_{m_1}$, and $\mathbf{e}_{m_2}$, as input data.
- The reverberant part of the impulse response is computed by subtracting the direct-path impulse response

$$\mathbf{r}(t;\boldsymbol{\theta}_m) = \mathbf{h}(t) - \mathbf{h}_d(t;\tau_0) \tag{13}$$

where the notation $(\boldsymbol{\theta}_m)$ was introduced to indicate that $\mathbf{r}(t;\boldsymbol{\theta}_m)$ is the mth realization.

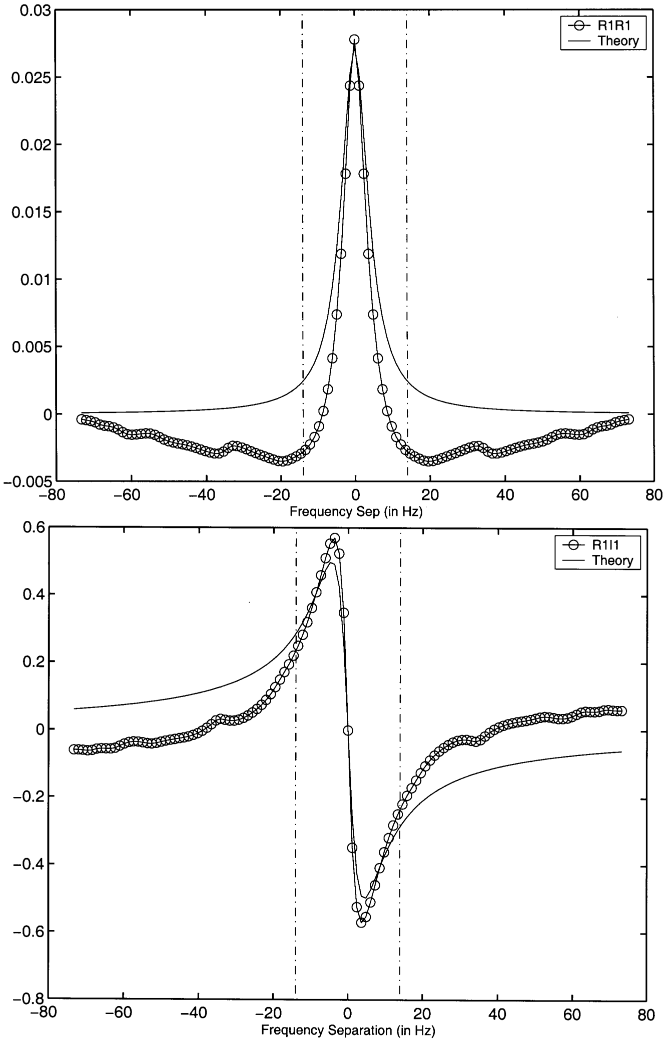

Fig. 1. Simulated data: Estimated correlation functions of $R_1(\omega;\boldsymbol{\theta})$. Here $r = 3$ m, the number of realizations of $R_1(\omega;\boldsymbol{\theta})$ equals 250, and $T_{60} = 0.5$ s. Vertical dash-dotted line indicates the coherence bandwidth $\rho = 7/T_{60}$, and the solid line illustrates the theoretical correlation functions (8)–(9) and (12). Top: $\hat{q}^2\hat{\varphi}_{11}^{rr}(\Delta\omega/2\pi)$. Bottom: $\hat{\varphi}_{11}^{ri}(\Delta\omega/2\pi)$.

In this manner, we compute new realizations of $\mathbf{r}(t;\boldsymbol{\theta}_m)$ without affecting the direct-path transfer function. The image-method in addition requires the reflection coefficient β as input data. For a given value of T_{60}, the reflection coefficient β is computed from *Eyring's* formula [12]

$$\beta = \exp\left\{-\frac{13.82}{\left(\frac{1}{L_x} + \frac{1}{L_y} + \frac{1}{L_z}\right) c T_{60}}\right\}. \tag{14}$$

The sampling frequency is chosen as $F_s = 1/T_s = 10$ kHz, and $\mathbf{h}(nT_s)$ is truncated to F_sT_{60} samples. For each realization, the reverberant part of the transfer function, $\mathbf{R}(\omega;\boldsymbol{\theta}_m)$, is computed in the interval [300, 5000] Hz.

For a particular outcome of $\mathbf{R}(\omega;\boldsymbol{\theta}_m)$, we now estimate the auto-correlation $\hat{q}^2\hat{\varphi}_{kk}^{rr}(\Delta\omega;\boldsymbol{\theta}_m)$ and the normalized cross-correlation $\hat{\varphi}_{kk}^{ri}(\Delta\omega;\boldsymbol{\theta}_m)$. The final estimates of the correlation functions are obtained by averaging the estimated correlation

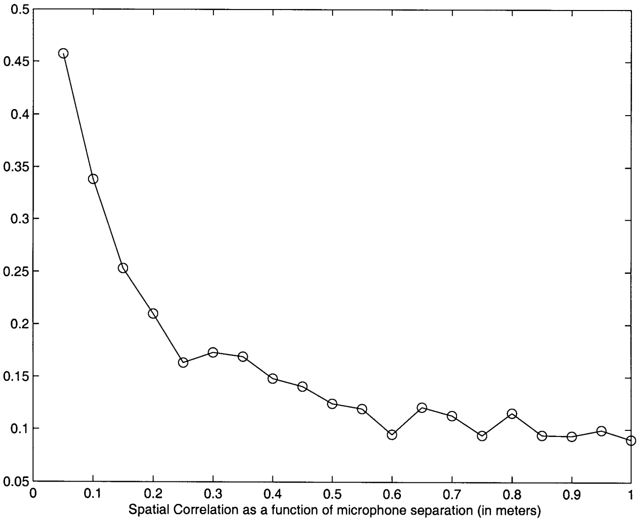

Fig. 2. Simulated data: Spatial (normalized) correlation $\hat{\varphi}_{12}^{rr}(\Delta\omega = 0)$, as a function of the microphone separation. Here, $r = 3$ m, the number of realizations of $\mathbf{R}(\omega;\boldsymbol{\theta})$ equals $250, T_{60} = 0.5$ s, and $\tau_0 = 0$.

functions using 250 random realizations of $\mathbf{R}(\omega;\boldsymbol{\theta})$. In this manner we simulate the averaging operator $E_{\boldsymbol{\theta}}\{\cdot\}$. In Fig. 1, the estimated as well as the theoretical (cf. (8)–(9)) correlation functions are illustrated for the case $T_{60} = 0.5$ s. The results in Fig. 1 indicate a good match between the theoretical and estimated correlation functions. In Fig. 2, we next plot the spatial correlation $\hat{\varphi}_{12}^{rr}(\Delta\omega = 0)$ as a function of d. As expected, $\hat{\varphi}_{12}^{rr}(\Delta\omega = 0)$ tends to decrease as the microphone separation d increases.

2) Real Data: Our next example deals with real data. At our disposal we had an office with dimensions $L_x = 6.6$ m, $L_y = 3.3$ m, and $L_z = 2.9$ m. The room was empty, except for a table in the middle of the room. A microphone was placed at $\mathbf{e}_m = [0.1\ 1.9\ 1.5]^T$. We next measured the transfer function from the microphone to 128 different source positions evenly distributed over the room. For all source positions, the loudspeaker was located at a fixed height of 1.1 m. The source signal was a chirp-signal (100–500 Hz). The microphone outputs were sampled with $F_s = 2000$ Hz, and the transfer functions were estimated at 1024 frequency points in the interval $[0, Fs/2]$ Hz.

Given the set of estimated transfer functions, correlation functions are estimated exactly as in the Section II-C1. However, let us first study the ensemble average of the *envelope* of the estimated impulse response, denoted $\hat{\bar{h}}(nT_s)$. For diffuse sound, $\hat{\bar{h}}(nT_s)$ should decay as $e^{-13.8nT_s/T_{60}}$, cf. [13]. Hence, $\log\{\hat{\bar{h}}(nT_s)\}$ should decay linearly as n increases. In Fig. 3, the outcome of $\log\{\hat{\bar{h}}(nT_s)\}$ is illustrated, and a straight-line (least squares) approximation of the reverberant part is illustrated as well. From the slope of the straight-line approximation we find that $T_{60} \simeq 0.35$ s, resulting in a Schroeder large room frequency $f_S \simeq 150$ Hz. Applying the estimated value of the reverberation time, we next computed the theoretical correlation functions φ_{11}^{rr} and φ_{11}^{ri}.

We remark that 1) according to *A3* the microphone is not located in the interior of the room and 2) the spectrum of $s(t)$ just barely satisfies the condition for the "Schroeder large room frequency." Despite this, Figs. 3 and 4 indicate a good agreement with the results of Section II-B. Note also that we did not eliminate the direct path part of the transfer function when

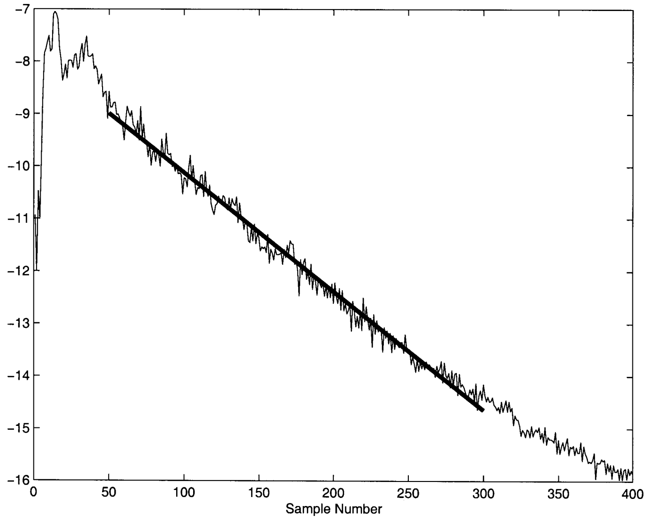

Fig. 3. Real data: ensemble average of the envelope of the impulse response, $\log\{\hat{\bar{h}}(nT_s)\}$.

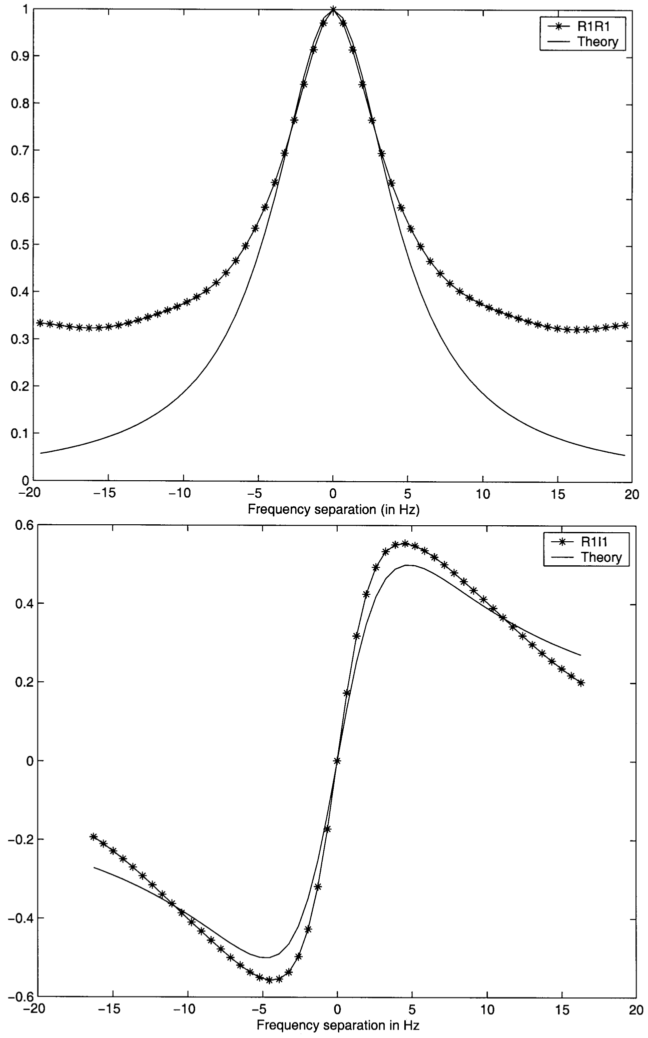

Fig. 4. Real data: estimated and theoretical (assuming that $T_{60} = 0.35$ s) correlation functions of the estimated room transfer function. Top: $\hat{\varphi}_{11}^{rr}(\Delta\omega/2\pi)$. Bottom: $\hat{\varphi}_{11}^{ri}(\Delta\omega/2\pi)$.

generating Figs. 3 and 4. This is most likely the reason why the theoretical auto-correlation in Fig. 4(a) decays more rapidly than the empirical one as the frequency separation increases. In Fig. 3, we further notice three different characteristics of $\log\{\hat{h}(nT_s)\}$: direct-path propagation for $5 \le n \le 15$, dominant early reflections for $15 \le n \le 40$, and late reverberation for $n > 40$. The main error in the introduced statistical model $\mathbf{h}(t) = \mathbf{h}_d(t) + \mathbf{r}(t)$ is then that the dominant early reflections are lumped together with the reverberant part.

III. Statistical Analysis

In this section we present a statistical analysis of the performance of source localization techniques in the presence of room reverberation. Using the statistical transfer function model developed in the previous section, the CRB for time-delay estimation and relevant Maximum Likelihood estimators are derived. In addition we analyze the performance of GCC, both in terms of asymptotic error variance and in terms of the probability of an anomalous estimate.

A. Single-Path Propagation TDE

Considering TDE for the single-path propagation model (1), the GCC [16] method is a popular alternative. GCC offers relatively high accuracy, and a modest computational complexity (FFT-based implementations). The estimated time-delay is obtained as

$$\hat{\tau} = \arg\max_{\tau} \hat{R}_{\mathrm{GCC}}(\tau) \tag{15}$$

where

$$\hat{R}_{\mathrm{GCC}}(\tau) \triangleq \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(\omega)|^2 \hat{P}_{12}(\omega) e^{j\omega\tau}\, d\omega. \tag{16}$$

Here, $\hat{P}_{12}(\omega) \triangleq T^{-1}X_2(\omega)X_1(\omega)^*$ denotes the estimated cross-power spectrum, T is the observation time, and $|G(\omega)|^2$ is a weighting function. The CRB (assuming $s(t), n_1(t)$ and $n_2(t)$ to be zero-mean, mutually uncorrelated wide-sense stationary Gaussian random processes with power spectra $P_{ss}(\omega), P_{n_1n_1}(\omega)$ and $P_{n_2n_2}(\omega)$) is further known to equal [16]

$$\mathrm{CRB}_{\mathrm{sp}}(\tau_0) = \left(2T\int_0^{\infty} \frac{\mathrm{SNR}(\omega)^2}{1+2\mathrm{SNR}(\omega)}\omega^2\, d\omega\right)^{-1}. \tag{17}$$

The subscript $(\cdot)_{\mathrm{sp}}$ indicates that the CRB is valid for the single-path propagation model (1). In (17), we introduced the signal to noise ratio $\mathrm{SNR}(\omega) = P_{ss}(\omega)/P_{nn}(\omega)$, where we for simplicity assumed that $P_{nn}(\omega) = P_{n_1n_1}(\omega) = P_{n_2n_2}(\omega)$. Note, with the above assumptions on $s(t), n_1(t)$ and $n_2(t)$, there exists a weighting $|G(\omega)|^2$ such that GCC is the Maximum Likelihood (ML) estimator of τ_0 [16].

B. Preliminaries

Since our main goal is to understand how room reverberation affects the performance of acoustical source localization methods, we will in the analysis typically assume that the additive measurement noise $\mathbf{n}(t)$ is absent. Given measurements of $\mathbf{x}(t)$ for $0 \le t \le T$, the frequency domain representation reads as

$$\begin{aligned}\mathbf{X}(\omega) = \begin{bmatrix} X_1(\omega) \\ X_2(\omega)\end{bmatrix} &= \int_0^T \mathbf{x}(t)e^{-j\omega t}\, dt \\ &= (\mathbf{H}_d(\omega;\tau_0) + \mathbf{R}(\omega;\boldsymbol{\theta}))S(\omega)\end{aligned} \tag{18}$$

where $S(\omega)$ denotes the Fourier transform of the source signal. Here, it is assumed that T is large so that the windowing distortion is negligible. To simplify the presentation, we scale the microphone outputs with $1/\kappa(r)$, and define the following quantities:

$$\begin{aligned}\mathbf{a}(t;\tau_0) &\triangleq \frac{1}{\kappa(r)}\mathbf{h}_d(t) \leftrightarrow \mathbf{A}(\omega;\tau_0) \\ &= \begin{bmatrix} 1 \\ e^{-j\omega\tau_0}\end{bmatrix}\end{aligned} \tag{19}$$

$$\mathbf{v}(t;\boldsymbol{\theta}) \triangleq \frac{1}{\kappa(r)}\mathbf{r}(t;\boldsymbol{\theta}) \leftrightarrow \mathbf{V}(\omega;\boldsymbol{\theta}) \tag{20}$$

where, consequently

$$E_{\boldsymbol{\theta}}\{\mathbf{V}(\omega;\boldsymbol{\theta})\mathbf{V}(\omega;\boldsymbol{\theta})^H\} = \underbrace{q^2/\kappa(r)^2}_{\sigma^2}\mathbf{I}_2 \tag{21}$$

and $(\cdot)^H$ denotes complex conjugate transpose. Assume that the source signal $s(t)$ is band-limited to the interval $[f_l, f_u]$ Hz, and that $f_l \ge f_S$. Suppose that the microphone output $\mathbf{x}(t)$ is sampled with sampling frequency $1/T_s$ Hz (assuming that $f_u \le 1/2T_s$), to produce the sequence $\mathbf{x}(nT_s), n = 0, 1, \ldots, N-1$, where $N = T/T_s$. For sampled data, (18) is computed using the Discrete Fourier Transform (DFT)

$$\mathbf{X}(\omega_k) = (\mathbf{A}(\omega_k;\tau_0) + \mathbf{V}(\omega_k;\boldsymbol{\theta}))S(\omega_k). \tag{22}$$

The DFT is computed for $\omega_k = 2\pi k/(NT_s), k = 0, 1, \ldots, N-1$. Due to the band-limited nature of $s(t)$, we consider $\mathbf{X}(\omega_k)$ only for $k = k_l, \ldots, k_u$, where $k_l = Nf_lT_s$ and $k_u = Nf_uT_s$.

C. CRB for TDE

We begin our statistical investigation by deriving the CRB for estimation of τ_0, based on the model (22). First we need to discuss how $S(\omega_k)$ should be modeled. One option is to assume that $\{S(\omega_k)\}_{k=k_l}^{k_u}$ are unknown but *deterministic* parameters. In the sensor array processing literature, it is well known that the corresponding CRB usually is too optimistic and hence unreachable, cf. [17].

In the following we instead assume that $s(t)$ is a zero mean wide-sense stationary random processes with power spectrum $P_{ss}(\omega)$. Conditioned on a fixed $\mathbf{V}(\omega_k;\boldsymbol{\theta})$, the DFT coefficients $S(\omega_k)$ are asymptotically (i.e., as $N \to \infty$) a sequence of uncorrelated zero-mean Gaussian random variables with variances $P_{ss}(\omega_k)$, see, e.g., [18, Ch. 15]. Imposing the natural assumption that $s(t)$ and $\mathbf{v}(t;\boldsymbol{\theta})$ are independent, it follows from the whiteness of $S(\omega_k)$ that $\mathbf{X}(\omega_k)$ and $\mathbf{X}(\omega_l)$ are uncorrelated (but not necessarily independent!). This observation holds true irrespective of the value of the coherence bandwidth ρ. To see this, study the expected value of the product of two arbitrary elements $\mathbf{X}(\omega_k)$ and $\mathbf{X}(\omega_l)$ $(k \neq l)$

$$\begin{aligned} &E_{\boldsymbol{\theta},S}\{\mathbf{X}(\omega_k)\mathbf{X}(\omega_l)^H\} \\ &\quad = E_S \underbrace{\{S(\omega_k)S(\omega_l)^*\}}_{=0} E_{\boldsymbol{\theta}}\{(\mathbf{A}(\omega_k;\tau_0) + \mathbf{V}(\omega_k;\boldsymbol{\theta})) \\ &\qquad \times (\mathbf{A}(\omega_l;\tau_0) + \mathbf{V}(\omega_l;\boldsymbol{\theta}))^H\} = 0. \end{aligned} \tag{23}$$

For the above result to hold true, the expectation operator must be defined as *the ensemble average over all signal realizations, and over all source/microphone positions*, denoted as $E_{\boldsymbol{\theta},S}\{\,\cdot\,\}$.

The remaing and most difficult issue is to determine the statistical distribution of $\mathbf{X}(\omega_k)$. For N large $S(\omega_k)$ is Gaussian, and from the central limit theorem one could argue that also $\mathbf{V}(\omega_k;\boldsymbol{\theta})$ is Gaussian. The random vector $\mathbf{X}(\omega_k)$ is then a sum of $\{\mathbf{A}(\omega_k;\tau_0)S(\omega_k)$, a Gaussian random vector$\}$ and $\{\mathbf{V}(\omega_k;\boldsymbol{\theta})S(\omega_k)$, a product of two uncorrelated Gaussian random vectors$\}$. It hence seems difficult to establish the distribution of $\mathbf{X}(\omega_k)$. In the following, we ignore that the distribution of $\mathbf{X}(\omega_k)$ is unknown, and perform the calculations pretending that $\mathbf{X}(\omega_k)$ is Gaussian.

Proposition 1: Suppose that $\{\mathbf{X}(\omega_k)\}_{k=k_l}^{k_u}$ is a sequence of uncorrelated zero-mean Gaussian random variables with $E_{\boldsymbol{\theta},S}\{\mathbf{X}(\omega_k)\mathbf{X}(\omega_k)^T\} = 0$ and

$$\begin{aligned} &E_{\boldsymbol{\theta},S}\{\mathbf{X}(\omega_k)\mathbf{X}(\omega_k)^H\} \\ &\qquad = P_{ss}(\omega_k)(\sigma^2\mathbf{I}_M + \mathbf{A}(\omega_k;\tau_0)\mathbf{A}(\omega_k;\tau_0)^H). \end{aligned} \tag{24}$$

Then, any unbiased estimator $\hat{\tau}$ of the true time-delay τ_0 satisfies

$$\begin{aligned} E_{\boldsymbol{\theta},S}\{(\hat{\tau} - \tau_0)^2\} &\geq \mathrm{CRB}_{\mathrm{rev}}(\tau_0) \\ &= \left(2\frac{\mathrm{SRR}^2}{1 + 2\mathrm{SRR}}\sum_{k=k_l}^{k_u}\omega_k^2\right)^{-1} \end{aligned} \tag{25}$$

where subscript "rev" indicates CRB for the room reverberation model, and the Signal to Reverberation Ratio (SRR) is defined as

$$\mathrm{SRR} \triangleq \frac{1}{\sigma^2} = \frac{\mathcal{A}(1-\beta^2)}{16\pi r^2\beta^2}. \tag{26}$$

Note, although the derivation is performed for the discrete time case, the dimension of (25) is $[\mathrm{seconds}]^2$.

Proof: Standard calculations which are omitted.

Based on empirical observations it was recently suggested that the CRB (17) for single-path propagation is valid also for reverberant environments, with the distinction that $\mathrm{SNR}(\omega)$ is modified to account for the effects of room reverberation. The "equivalent SNR" suggested in [8] reads as

$$(\mathrm{SNR}_{\mathrm{eq}}(\omega))_i = \frac{|H_i(\omega;0)|^2P_{ss}(\omega)}{|H_i(\omega;\beta) - H_i(\omega;0)|^2P_{ss}(\omega) + P_{n_in_i}(\omega)} \tag{27}$$

where $H_i(\omega;0)$ denotes the transfer function from the source to the ith microphone in case of no reverberation, and $H_i(\omega;\beta)$ denotes the same transfer function with reverberation included. In order to relate Proposition 1 to the findings in [8], let us include the effects of additive measurement noise in the CRB expression (25). Assuming that the additive noise $\mathbf{n}(nT_s)$ is zero-mean, white, Gaussian, and with variance $E\{\mathbf{n}(nT_s)\mathbf{n}(nT_s)^T\} = \eta^2\mathbf{I}_2$, the CRB expression (25) for the case with additive measurement noise present reads as

$$\overline{\mathrm{CRB}}_{\mathrm{rev}}(\tau_0) = \left(2\sum_{k=k_l}^{k_u}\frac{\mathrm{SNRR}(\omega_k)^2}{1 + 2\mathrm{SNRR}(\omega_k)}\omega_k^2\right)^{-1} \tag{28}$$

where the signal to noise and reverberation ratio (SNRR) is defined as

$$\mathrm{SNRR}(\omega) = \frac{P_{ss}(\omega)\frac{1}{4\pi r^2}}{P_{ss}(\omega)\frac{4\beta^2}{\mathcal{A}(1-\beta^2)} + \eta^2}. \tag{29}$$

Note that $P_{ss}(\omega)$ appears in (28), in contrast to the expression (25).

Proposition 1 can then be related to the findings in [8] by observing that SNRR in (29) corresponds to the average value of $\mathrm{SNR}_{\mathrm{eq}}$ (27), assuming that

1) the quantity $(H_i(\omega;\beta) - H_i(\omega;0))$ corresponds to diffuse sound;
2) $d \ll (\|\mathbf{e}_s - \mathbf{e}_{m_1}\|, \|\mathbf{e}_s - \mathbf{e}_{m_2}\|)$ so that the two microphones receive an equal amount of energy from the direct path.

D. Analysis of GCC

The purpose of the following section is derive the variance of $\hat{\tau}$ as defined in (15). Analysis of $\hat{\tau}$ is in general difficult due to the combined effects of a finite measurement time and the nondeterministic nature of $\mathbf{V}(\omega;\boldsymbol{\theta})$. To simplify the statistical analysis we will neglect the influence of a finite observation time, and focus on the effects of room reverberation.

Using $E_S\{\hat{P}_{12}(\omega)\}$ instead of $\hat{P}_{12}(\omega)$ in (16), we find that $\hat{\tau} = \arg\max_\tau Q(\tau)$ where $Q(\tau) \triangleq E_S\{\hat{R}_{\mathrm{GCC}}(\tau)\}$ and

$$Q(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}|G(\omega)|^2P_{ss}(\omega)P_\epsilon(\omega)e^{j\omega\tau}\,d\omega \tag{30}$$

$$\begin{aligned} P_\epsilon(\omega) = e^{-j\omega\tau_0} + &\underbrace{V_2(\omega) + V_1(\omega)^*e^{-j\omega\tau_0}}_{\epsilon(\omega)} \\ &+ \underbrace{V_2(\omega)V_1(\omega)^*}_{\tilde{\epsilon}(\omega)}. \end{aligned} \tag{31}$$

To simplify the notation, argument $(\cdot;\boldsymbol{\theta})$ has been suppressed. To perform the analysis, we assume that $\hat{\tau}$ is consistent in the sense that $\hat{\tau} \to \tau_0$ as $\sigma \to 0$. We will not discuss precise conditions for this to hold, but simply assume that the user has "chosen" d and $s(t)$ so that $\hat{\tau} = \tau_0$ in the absence of reverberation.

The basic idea in the following is to compute the variance of $\hat{\tau}$ for large SRR's. Since $\hat{\tau}$ maximizes $Q(\tau), Q'(\hat{\tau}) = 0$, where $Q'(\hat{\tau})$ denotes the gradient of $Q(\tau)$ evaluated at $\hat{\tau}$. For SRR large, a first order Taylor expansion yields

$$0 \simeq Q'(\tau_0) + Q''(\tau_0)(\hat{\tau} - \tau_0) \tag{32}$$

where $Q''(\tau_0)$ denotes the Hessian. It now follows that

$$\hat{\tau} - \tau_0 \simeq -\frac{1}{Z}Q'(\tau_0) \tag{33}$$

where $Z \triangleq \lim_{\sigma \to 0} Q''(\tau_0)$. For high SRR the variance of $\hat{\tau}$ is thus given by

$$E_{\boldsymbol{\theta}}\{(\hat{\tau} - \tau_0)^2\} = \sigma^2 \frac{K}{Z^2} + o(\sigma^2) \tag{34}$$

where

$$K \triangleq \lim_{\sigma \to 0} \frac{1}{\sigma^2} E_{\boldsymbol{\theta}}\{(Q'(\tau_0))^2\}. \tag{35}$$

Assuming that $V_1(\omega)$ and $V_2(\omega)$ are Gaussian random variables with a negligible spatial correlation it follows that $E_{\boldsymbol{\theta}}\{\epsilon(\omega)\epsilon(\omega)^*\} = 2\sigma^2, E_{\boldsymbol{\theta}}\{\tilde{\epsilon}(\omega)\tilde{\epsilon}(\omega)^*\} = \sigma^4$, and $E_{\boldsymbol{\theta}}\{\epsilon(\omega)\tilde{\epsilon}(\omega)^*\} = 0$. From these relationships, and since we in (34) neglect all terms that are of order $o(\sigma^2), P_\epsilon(\omega)$ can be approximated as

$$P_\epsilon(\omega) \simeq e^{-j\omega\tau_0} + \epsilon(\omega). \tag{36}$$

Applying the approximation (36), the gradient $Q'(\tau_0)$ can be written as

$$Q'(\tau_0) = \frac{1}{2\pi} \int_{\omega_l}^{\omega_u} |G(\omega)|^2 P_{ss}(\omega) j\omega (\epsilon(\omega) e^{j\omega\tau_0} - \epsilon(\omega)^* e^{-j\omega\tau_0})\, d\omega. \tag{37}$$

Consider then the computation of K (35). The main technical difficulty in computing K is the fact that $\epsilon(\omega_1)$ and $\epsilon(\omega_2)$ have a nonnegligible correlation, unless $|\omega_1 - \omega_2| > 2\pi\rho$. Although this correlation with some effort can be included in the calculations, we have chosen a simpler but approximate approach. Introduce the following sampling of the frequency axis:

$$\phi_k = \omega_l + k \cdot \delta_\omega, \quad k = 0, \dots, L. \tag{38}$$

The sampling interval is in principle defined as $\delta_\omega = 2\pi\rho$, but a small perturbation to $2\pi\rho$ is allowed to ensure that L is an integer (i.e., so that $\phi_L = \omega_u$). This perturbation will only have a small effect on the final result.

For $k \neq l$, we can then assume that $\epsilon(\phi_k)$ and $\epsilon(\phi_l)$ are uncorrelated random variables. In the following we also require that $P_{ss}(\omega)$ (and $|G(\omega)|^2$) can be considered constant within each frequency interval $[\phi_k, \phi_{k+1})$. That is, for $\tilde{\omega} \in [\phi_k, \phi_{k+1})$, it is assumed that $P_{ss}(\tilde{\omega}) = P_{ss}(\phi_k)$ (and that $|G(\tilde{\omega})|^2 = |G(\phi_k)|^2$). The gradient $Q'(\tau_0)$ can then be approximated in the following manner:

$$Q'(\tau_0) \simeq \frac{j}{2\pi} \sum_{k=0}^{L-1} |G(\phi_k)|^2 P_{ss}(\phi_k)(\epsilon(\phi_k)\Gamma_k - \epsilon(\phi_k)^* \Gamma_k^*) \tag{39}$$

where

$$\Gamma_k \triangleq \int_{\phi_k}^{\phi_{k+1}} \omega e^{j\omega\tau_0}\, d\omega = -\frac{1}{\tau_0^2}(e^{j\phi_{k+1}\tau_0}(j\phi_{k+1}\tau_0 - 1) - e^{j\phi_k\tau_0}(j\phi_k\tau_0 - 1)). \tag{40}$$

For $\delta_\omega \tau_0 \ll 1$ (i.e., $2\pi\rho d/c \ll 1$), $e^{j\phi_{k+1}\tau_0} = e^{j\phi_k\tau_0 + j\delta_\omega\tau_0} \simeq e^{j\phi_k\tau_0}(1 + j\delta_\omega\tau_0)$, and Γ_k can be simplified as

$$\Gamma_k \simeq e^{j\phi_k\tau_0}((\delta_\omega)^2 + \delta_\omega\phi_k) \simeq e^{j\phi_k\tau_0}\delta_\omega\phi_k \tag{41}$$

where the last approximation is applicable as long as $\delta_\omega \ll \omega_k$, i.e., $T_{60} \gg 7/f_l$ ($\simeq$0.02 s for speech). The final approximation of the gradient then reads as

$$Q'(\tau_0) \simeq \frac{j \cdot \delta_\omega}{2\pi} \sum_{k=0}^{L-1} |G(\phi_k)|^2 P_{ss}(\phi_k)\phi_k(\epsilon(\phi_k)e^{j\phi_k\tau_0} - \epsilon(\phi_k)^* e^{-j\phi_k\tau_0}). \tag{42}$$

Since $\epsilon(\phi_k)$ and $\epsilon(\phi_l)$ are assumed uncorrelated for $k \neq l$, it is now straightforward to compute K

$$K = \frac{(\delta_\omega)^2}{\pi^2} \sum_{k=0}^{L-1} |G(\phi_k)|^4 P_{ss}(\phi_k)^2 \phi_k^2. \tag{43}$$

Since $P_\epsilon(\omega) \to e^{-j\omega\tau_0}$ as $\sigma \to 0$, the Hessian evaluated at $\tau = \tau_0$, can be written as

$$Z = -\frac{1}{\pi} \int_{\omega_l}^{\omega_u} |G(\omega)|^2 P_{ss}(\omega)\omega^2\, d\omega \tag{44}$$

or with approximations similar to the ones that lead to (42)

$$Z \simeq -\frac{\delta_\omega}{\pi} \sum_{k=0}^{L-1} |G(\phi_k)|^2 P_{ss}(\phi_k)\phi_k^2. \tag{45}$$

Using (43) and (45), the large SRR mean square error (34) can now be evaluated:

$$E_{\boldsymbol{\theta}}\{(\hat{\tau} - \tau_0)^2\} = \sigma^2 \frac{\sum_{k=0}^{L-1} |G(\phi_k)|^4 P_{ss}(\phi_k)^2 \phi_k^2}{\left(\sum_{k=0}^{L-1} |G(\phi_k)|^2 P_{ss}(\phi_k)\phi_k^2\right)^2} + o(\sigma^2). \tag{46}$$

The mean square error (46) clearly depends on how the weightings $|G(\phi_k)|^2$ are chosen. An interesting question is how $|G(\phi_k)|^2$ should be chosen for lowest possible error variance.

Proposition 2: The error variance $\sigma^2 K/Z^2$ is minimized if

$$|G(\phi_k)|^2 = \frac{1}{P_{ss}(\phi_k)}. \tag{47}$$

Proof: Define the following matrices:

$$\boldsymbol{\Delta}^T \triangleq [\phi_0\sqrt{P_{ss}(\phi_0)} \quad \cdots \quad \phi_{L-1}\sqrt{P_{ss}(\phi_{L-1})}] \tag{48}$$

$$\mathbf{G} \triangleq \mathrm{diag}\{|G(\phi_0)|^2, \dots, |G(\phi_{L-1})|^2\} \tag{49}$$

$$\boldsymbol{\Sigma} \triangleq \mathrm{diag}\{P_{ss}(\phi_0), \dots, P_{ss}(\phi_{L-1})\}. \tag{50}$$

Then, it is easy to see that

$$\frac{K}{Z^2} = (\boldsymbol{\Delta}^T\mathbf{G}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^T\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}\boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{G}\boldsymbol{\Delta})^{-1}. \tag{51}$$

Assuming that $P_{ss}(\phi_k) > 0$ for $k = 0, \dots, L-1$, it follows from well-known matrix optimization results (see e.g., ([19], Appendix II.2)) that the best possible weightings are given by $|G(\phi_k)|^2 = 1/P_{ss}(\phi_k)$.

In general, the quantity $P_{ss}(\phi_k)$ is unknown. However, note that

$$|E_{\boldsymbol{\theta},S}\{\hat{P}_{12}(\phi_k)\}| = |P_{ss}(\phi_k)e^{-j\phi_k\tau_0}| = P_{ss}(\phi_k). \tag{52}$$

Hence, a natural estimate of $P_{ss}(\phi_k)$ is $\hat{P}_{ss}(\phi_k) = |\hat{P}_{12}(\phi_k)|$, which interestingly enough results in a GCC-method known as the PHAse Transform (PHAT) time-delay estimator [16]. The above calculations then indicate that PHAT is the GCC method most suitable for reverberant environments, which agrees well with earlier observations cf. [20]–[22].

For large SRRs, our numerical experience is that the derived results quite accurately predict the actual performance. As the SRR decreases, we have however observed a discrepancy due to the fact that terms of order σ^4 have been neglected. Including the effects of $\tilde{\epsilon}(\omega)$ [cf. (31)], the GCC error variance with respect to $|G(\phi_k)|^2 = 1/P_{ss}(\phi_k)$ reads as

$$E_{\boldsymbol{\theta}}\{(\hat{\tau} - \tau_0)^2\} \geq \left(2\frac{\mathrm{SRR}^2}{1 + 2\mathrm{SRR}} \sum_{k=k_l}^{k_u} \phi_k^2\right)^{-1}. \quad (53)$$

Note that the dimension of (53) is $[\mathrm{seconds}]^2$, and observe the similarity with the CRB (25).

E. Probability of an Anomalous Estimate

Knowledge of the large SRR mean square error (53) for a particular room configuration is certainly an important piece of information. However, expression (53) is rather local in the sense that it is reachable only for large SRRs.

In practical applications, the limiting factor of the performance is typically the fact that GCC-based localization methods suffer from outliers, simply because the "wrong peak" of the GCC function $\hat{R}_{\mathrm{GCC}}(\tau)$ is selected. It is therefore of interest to analyze the probability of an anomalous estimate. For the case with single-path propagation and additive uncorrelated measurement noise, Ianniello analyzed this probability in a classical paper [23]. The purpose of the following section is then to extend Ianniello's analysis to include the effects of room reverberation.

Analysis of the outlier probability is difficult, and as in the previous section we have to make a couple of simplifying assumptions. To begin with, let us assume that the zero-mean wide-sense stationary random process $s(t)$ fulfills

$$P_{ss}(\omega) = \begin{cases} P_0 & |\omega| \leq \omega_u \\ 0 & \text{else} \end{cases}. \quad (54)$$

We further assume that the GCC weighting equals $|G(\omega)|^2 = 1$. As in Section III-D, we study only the behavior of the random variable $Q(\tau)$.

When the power spectrum of $s(t)$ satisfies (54)

$$E_{\boldsymbol{\theta}}\{Q(\tau)\} = \frac{1}{\pi} P_0 \omega_u \operatorname{sinc}(\omega_u(\tau - \tau_0)). \quad (55)$$

Note that $E_{\boldsymbol{\theta}}\{Q(\tau_0)\} = P_0\omega_u/\pi$, and $E_{\boldsymbol{\theta}}\{Q(\tau_0 + (l\pi)/\omega_u)\} = 0$ for integers $l = \pm 1, \pm 2, \cdots$. Define next the maximum possible time-delay as $\tau_M = d/c$. Let l_1 be the smallest integer that fulfills $\tau_0 + l_1\pi/\omega_u \geq -\tau_M$, and let l_2 be the largest integer that fulfills $\tau_0 + l_2\pi/\omega_u \leq \tau_M$. Define also the $(l_2 - l_1 + 1)$-dimensional vector $\mathbf{z}$ from an equidistant sampling of $Q(\tau)$

$$\mathbf{z} = [Q(\tau_0 + l_1\pi/\omega_u), \ldots, Q(\tau_0), \ldots, Q(\tau_0 + l_2\pi/\omega_u)]. \quad (56)$$

Exactly as in [23] we have used (the inverse of) the band-width of $s(t)$ to sample $Q(\tau)$ with sample interval π/ω_u, and as in [23] an anomalous event is defined as

$$\mathcal{E} \triangleq [Q(\tau_0 + l\pi/\omega_u) > Q(\tau_0) \text{ for at least one } l \in \mathcal{L}] \quad (57)$$

where $\mathcal{L} \triangleq \{l_1, \ldots, -1, 1, \ldots, l_2\}$. To be able to compute $\mathrm{Prob}[\mathcal{E}]$ we require knowledge of the statistical properties of the vector $\mathbf{z}$. Once again we will encounter difficulties incorporating the effects of the correlation-function of $\mathbf{V}(\omega; \boldsymbol{\theta})$, and as in the previous section we simplify the problem with the following approximation:

$$Q(\tau) \simeq \bar{Q}(\tau) \triangleq \frac{\delta_\omega}{2\pi} \sum_{k=1}^{L} P_0 \left(e^{j(\tau - \tau_0)\phi_k} + \varepsilon(\phi_k) e^{j\phi_k \tau}\right) \quad (58)$$

where $\phi_k = -\omega_u + k \cdot \delta_\omega, L = 2\omega_u/\delta_\omega$, and

$$\varepsilon(\omega) = V_2(\omega) + V_1(\omega)^* e^{-j\omega\tau_0} + V_2(\omega)V_1(\omega)^*. \quad (59)$$

The frequency spacing δ_ω is defined as in Section III.D, i.e., $\delta_\omega = 2\pi\rho$. Define the vector $\bar{\mathbf{z}}$ as

$$\bar{\mathbf{z}} = [\bar{Q}(\tau_0 + l_1\pi/\omega_u), \ldots, \bar{Q}(\tau_0), \ldots, \bar{Q}(\tau_0 + l_2\pi/\omega_u)]. \quad (60)$$

Consider the expected value of an arbitrary element of $\bar{\mathbf{z}}$. Since $E_{\boldsymbol{\theta}}\{\varepsilon(\omega)\} = 0$

$$\begin{aligned} E_{\boldsymbol{\theta}}\{\bar{Q}(\tau_0 + l\pi/\omega_u)\} &= \frac{\delta_\omega}{2\pi} \sum_{k=1}^{L} P_0 e^{j\frac{l\pi}{\omega_u}\phi_k} \\ &= \frac{\delta_\omega P_0}{2\pi} e^{-jl\pi} \sum_{k=1}^{L} e^{\frac{j2\pi k}{L} l} \\ &= \begin{cases} P_0\omega_u/\pi & l = 0 \\ 0 & \text{else} \end{cases}. \end{aligned} \quad (61)$$

Hence, $E_{\boldsymbol{\theta}}\{\bar{\mathbf{z}}\} = E_{\boldsymbol{\theta}}\{\mathbf{z}\}$. Consider next the covariance between two arbitrary elements of $\bar{\mathbf{z}}$

$$\begin{aligned} &\mathrm{Cov}_{\boldsymbol{\theta}}\{\bar{Q}(\tau_0 + m\pi/\omega_u), \bar{Q}(\tau_0 + n\pi/\omega_u)\} \\ &\quad = \frac{\delta_\omega^2 P_0^2}{4\pi^2}(2\sigma^2 + \sigma^4) e^{-j\pi(m-n)} \sum_{k=1}^{L} e^{\frac{j2\pi k}{L}(m-n)} \\ &\quad = \begin{cases} \frac{P_0^2\omega_u^2}{\pi^2 L}(2\sigma^2 + \sigma^4) & m = n \\ 0 & m \neq n \end{cases}. \end{aligned} \quad (62)$$

Hence, the components of $\bar{\mathbf{z}}$ are uncorrelated with identical variances. The remaining issue is to find the distribution of $\bar{\mathbf{z}}$. For finite L the distribution of $\bar{\mathbf{z}}$ is difficult to find, simply since we do not know the distribution of $\varepsilon(\omega)$. For L large, on the other hand, we invoke the central limit theorem to argue that $\bar{\mathbf{z}}$ is Gaussian. Note that the factor $P_0\omega_u/\pi$, which appears in (61) and (62), does not affect the computation of $\mathrm{Prob}[\mathcal{E}]$ and can be replaced with e.g., unity. It is thus the quantities $\alpha^2 \triangleq L^{-1}(2\sigma^2 + \sigma^4)$ and $l_2 - l_1$ that determines the outlier probability. Assuming that $\bar{\mathbf{z}}$ is Gaussian, and that $\bar{\mathbf{z}} \simeq \mathbf{z}$, $\mathrm{Prob}[\mathcal{E}]$ can be evaluated as [23]

$$\mathrm{Prob}[\mathcal{E}] = 1 - \int_{-\infty}^{\infty} p(z_0) \left\{\int_{-\infty}^{z_0} p(z_l)\, dz_l\right\}^{l_2 - l_1} dz_0. \quad (63)$$

Here, we introduced the following notation.

- $z_0 \triangleq Q(\tau_0)$, i.e., the GCC function evaluated at the true time-delay. With the above assumptions, $z_0 \in \mathcal{N}(1, \alpha^2)$.
- $z_l \triangleq Q(\tau_0 + l\pi/\omega_u), n \in \mathcal{L}$. With the above assumptions, $z_l \in \mathcal{N}(0, \alpha^2)$.
- $p(z_0)$: PDF of z_0.
- $p(z_l)$: PDF of z_l.

The integral (63) typically has to be evaluated using numerical integration.

In a practical setup with $s(t)$ representing speech, assumption (54) is typically violated. However, we conjecture that the above results are relevant also for colored source signals, assuming that we apply a GCC method (such as PHAT) that employs pre-whitening. We will return to this issue in the numerical examples.

IV. ML Estimation of τ_0

Although we have shown that PHAT may be considered as the GCC-estimator most suitable for reverberant environments, it may be of interest to study the ML estimator of τ_0 for the model (22). It should also be noted that PHAT is applicable only to the case with two microphones, whereas the ML estimators to be proposed easily can include measurements from $M \geq 2$ microphones.

Assume that the sequence $\{\mathbf{X}(\omega_k\}_{k=k_l}^{k_u}$ consists of uncorrelated and Gaussian random variables

$$\mathbf{X}(\omega_k) \in \mathcal{N}(0, P_{ss}(\omega_k)(\sigma^2\mathbf{I}_M + \mathbf{A}(\omega_k;\tau_0)\mathbf{A}(\omega_k;\tau_0)^H)). \tag{64}$$

Here, M denotes the number of microphones, i.e., $M = 2$. Concentrating the log-likelihood function with respect to $\{P_{ss}(\omega_k)\}_{k=k_l}^{k_u}$, straightforward calculations show that the ML estimator of τ_0 and σ^2 is obtained by minimizing the following criterion function:

$$V_{\text{SML}}(\tau, \sigma^2) = (k_u - k_l + 1)\left(1 + \frac{M}{\sigma^2}\right) + M\sum_{k=k_l}^{k_u} \log\left\{1 - \frac{\|\mathbf{X}(\omega_k)^H\mathbf{A}(\omega_k;\tau)\|^2}{(\sigma^2 + M)\|\mathbf{X}(\omega_k)\|^2}\right\}. \tag{65}$$

Unfortunately, $V_{\text{SML}}(\cdot)$ cannot be concentrated with respect to σ^2, which leads to a prohibitively large computational complexity. Therefore, we consider the resulting estimator to be of theoretical interest only.

We next derive an Approximate ML (AML) estimator, which is more attractive from a computational point of view. Assume that $\{\mathbf{X}(\omega_k)\}_{k=k_l}^{k_u}$ consists of uncorrelated and Gaussian random variables with the following distribution:

$$\mathbf{X}(\omega_k) \in \mathcal{N}(\mathbf{A}(\omega_k)S(\omega_k), \sigma_k^2\mathbf{I}_2). \tag{66}$$

Hence, $\{S(\omega_k)\}_{k=k_l}^{k_u}$ are modeled as unknown but *deterministic* parameters. We further ignore the fact that the variance of $\mathbf{X}(\omega_k)$ is proportional to $|S(\omega_k)|^2$. Instead, we allow the parameters σ_k^2 to be frequency dependent.

Concentrating the resulting likelihood function with respect to "nuisance" parameters, the following criterion function is obtained:

$$V_{\text{AML}}(\tau) = \sum_{k=k_l}^{k_u} \log\{\|\mathbf{\Pi}_k^{\perp}(\tau)\mathbf{X}(\omega_k)\|^2\} \tag{67}$$

where $\mathbf{\Pi}_k^{\perp}(\tau)$ denotes the projection matrix onto the orthogonal complement of the space spanned by $\mathbf{A}(\omega_k;\tau)$

$$\mathbf{\Pi}_k^{\perp}(\tau) \triangleq \mathbf{I}_M - \frac{\mathbf{A}(\omega_k;\tau)\mathbf{A}(\omega_k;\tau)^H}{M}. \tag{68}$$

The time-delay is estimated by minimizing $V_{\text{AML}}(\tau)$ with respect to τ. The criterion function $V_{\text{AML}}(\tau)$ is nonlinear in τ, but in contrast to $V_{\text{SML}}(\cdot)$ it does not depend on σ^2.

V. Numerical Examples

A. Properties of $Q(\tau)$

In the first examples, we will assume that an infinite number of measurements of $\mathbf{x}(t)$ is available. The derived CRB is then not relevant, and we focus on the outlier probability (63) and on the GCC error variance (53).

We study a scenario identical to the one in Section II-C1, and we generate realizations of the room transfer function in an identical manner. We assume that the sampled source signal $s(nT_s)$ is white, and that $1/T_s = 10$ kHz. The source signal then satisfies (54), with $\omega_u/(2\pi) = 5$ kHz.

For each realization of $\mathbf{h}(t)$, we compute $Q(\tau)$ for $\tau^l \triangleq \tau_0 + l\pi/\omega_u, l \in [0, \mathcal{L}]$, and τ_0 is estimated from the maximum of $Q(\tau^l)$. To get subsample resolution, the estimated time-delay is refined using quadratic interpolation, cf. [23].

Suppose that $r = 3$ m, and vary the reverberation time in the interval [0.03, 1] s. Since $s(nT_s)$ is white, also frequencies that are below the Schroeder large room frequency excites the room transfer function. However, for the studied interval of reverberation times, $f_S \in [24, 143]$ Hz. The fraction of the signal spectrum not satisfying condition *A2* is hence considered small.

Let us study the performance of GCC, assuming that $|G(\omega)|^2 = 1$. In the computation of the GCC error variance (53) and the outlier probability (63) we have used $\rho = 10/T_{60}$ rather than the "derived" expression (10). Note, in agreement with the definition in Section III-E, an outlier is defined as the following event; $Q(\tau_0) < Q(\tau_0 + lT_s)$ for at least one $l \in \mathcal{L}$. To present the empirical standard deviation of $\hat{\tau}$ in a fair manner, we exclude estimates not satisfying $|\hat{\tau} - \tau_0| < T_s$.

Consider then Fig. 5, which show a good match between the theoretical and the empirical results. We however note that the GCC error variance (53) is overly pessimistic for small values of T_{60} ($T_{60} < 0.07$ s). The reason is probably that the number of received echos is too few for diffuse sound to be present, invalidating the assumption that $\mathbf{r}(t)$ corresponds to diffuse sound propagation.

In Fig. 6, we next consider the influence of the microphone separation d. The simulation is identical to the previous one, with the distinction that we let $T_{60} = 0.4$ s and instead vary d in the interval [0.05, 1] m. For $d > 0.2$ m, the assumption that the spatial correlation of $\mathbf{V}(\omega;\boldsymbol{\theta})$ is negligible seems applicable.

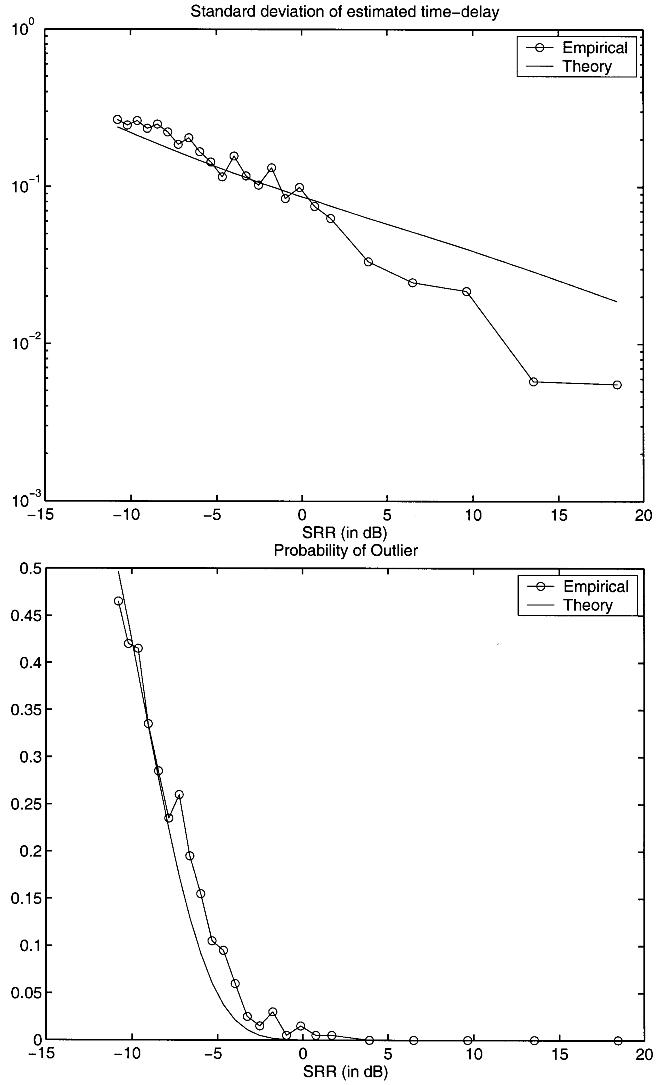

Fig. 5. Standard deviation and outlier probability of the estimated time-delay as a function of SRR. Here $r = 3$ m, $\tau_0 = 0, \rho = 10/T_{60}$, and the results are based on 300 Monte Carlo simulations.

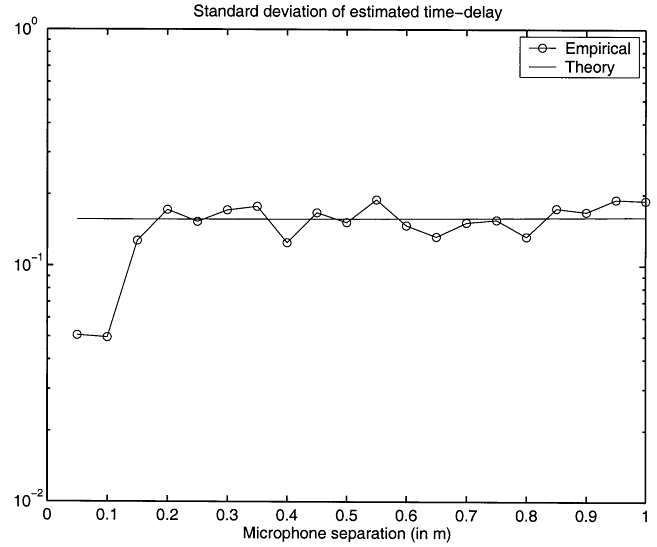

Fig. 6. Theoretical and empirical standard deviation of the estimated time-delay as a function of the microphone separation. Here $T_{60} = 0.4$ s, $r = 3$ m, $\rho = 10/T_{60}$, and the results are based on 300 Monte Carlo simulations.

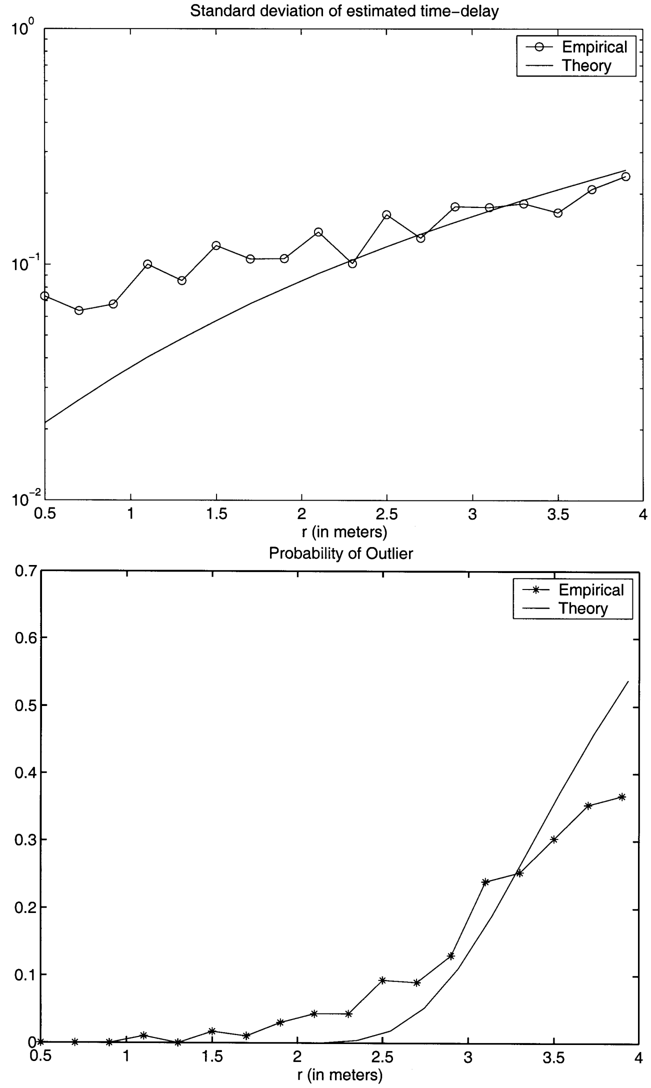

Fig. 7. Standard deviation and outlier probability of the estimated time-delay as a function of the distance between the source and the microphones. Here $d = 1$ m, $\tau_0 = 0, T_{60} = 0.4$ s, and the results are based on 300 Monte Carlo simulations.

Consider finally Fig. 7, where we study the GCC error variance for various distances between the source and the microphones. In Fig. 7 we applied a fixed $T_{60} = 0.4$ s, and $d = 1$ m. The empirical results show a good agreement with the theoretical error variance only for $r > 1.5$–2 m. A plausible explanation is that the results in Fig. 7 are affected by early reflections. When r decreases, the microphones will receive a number of strong early reflections, invalidating the assumption that $\mathbf{r}(t)$ corresponds to diffuse sound propagation. Hence, to fully explain the results in Fig. 7, one probably has to include the effect of dominant early reflections in the analysis.

B. Finite Sample Effects

In Fig. 8 we study the performance for a number of different sample sizes. The reverberation time is fixed at $T_{60} = 0.4$ s. We study the AML method together with two variants of GCC: PHAT and a variant with $|G(\omega)|^2 = 1$. In the simulation we have included additive white zero mean Gaussian measurement noise. The noise level is chosen such that the ratio between the

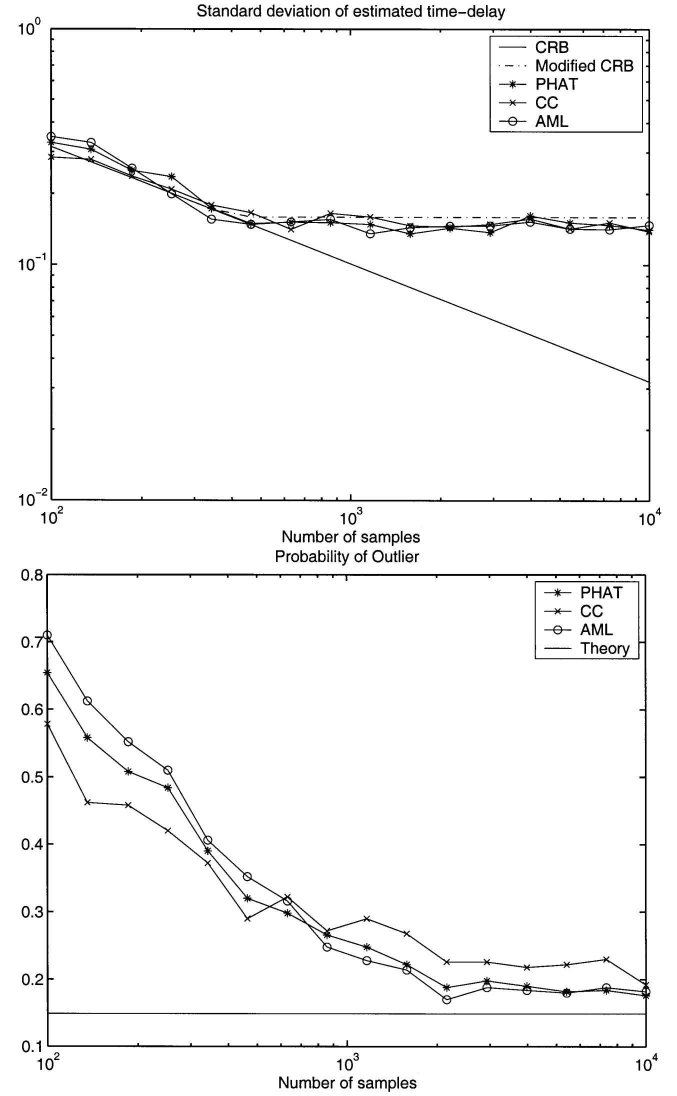

Fig. 8. Standard deviation and outlier probability of the estimated time-delay as a function of N. The source signal is a white Gaussian random process. Here, $T_{60} = 0.4$ s, $r = 3$ m, $\tau_0 = 0, \rho = 10/T_{60}$, and the results are based on 500 Monte Carlo simulations. Legend "CC" refers to the GCC method with $|G(\omega)|^2 = 1$.

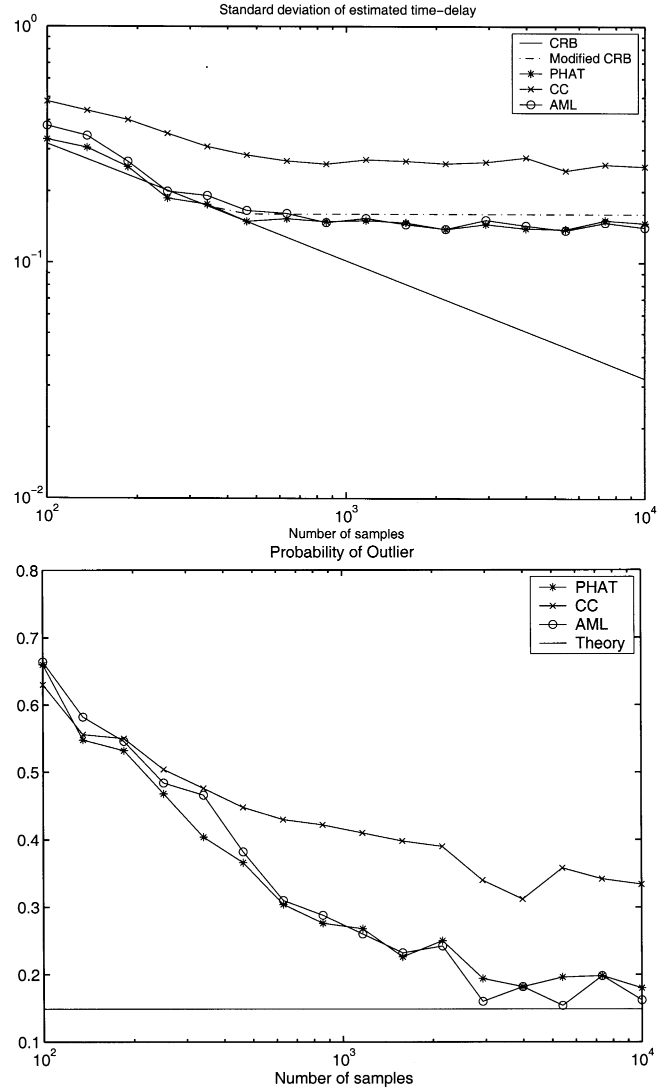

Fig. 9. Standard deviation and outlier probability of the estimated time-delay as a function of N. The source signal is generated as in (69). Here, $T_{60} = 0.4$ s, $r = 3$ m, $\tau_0 = 0, \rho = 10/T_{60}$, and the results are based on 500 Monte Carlo simulations. Legend "CC" refers to the GCC method with $|G(\omega)|^2 = 1$.

direct-path signal power and the noise power equals 40 dB. In Fig. 8 the source signal is a white Gaussian random process, whereas Fig. 9 illustrates the case with a colored source signal

$$s(n) - s(n-1) + 0.5s(n-2) = e(n) \tag{69}$$

where $e(n)$ is a white zero mean Gaussian random process. Since we would like to simulate the averaging operator $E_{\boldsymbol{\theta},S}\{\cdot\}$, we generate a new realization of the source signal for each realization of the room transfer function.

First, we notice that the derived CRB is reachable only for N small. This is most likely an effect of an (quite unrealistic) assumption on the Gaussianity of $\mathbf{X}(\omega)$. Based on our previous discussions, and assuming a realistic T_{60}, it is reasonable to assume that both $\mathbf{V}(\omega;\boldsymbol{\theta})$ and $S(\omega)$ are Gaussian. The product $\mathbf{V}(\omega;\boldsymbol{\theta})S(\omega)$ is then *not* Gaussian. The fact that $\mathbf{X}(\omega_k)$ and $\mathbf{X}(\omega_l)$ are uncorrelated for $k \neq l$, then does not imply that $\mathbf{X}(\omega_k)$ and $\mathbf{X}(\omega_l)$ are independent and the frequency correlation of $\mathbf{V}(\omega;\boldsymbol{\theta})$ will affect the performance. In Figs. 8–9 we have handled this issue in a heuristic manner (legend "Modified CRB"). As previously discussed, the reverberant transfer function has a coherence bandwidth ρ. Within the band-width of the source signal there are approximately $\tilde{N} \triangleq 2(f_u - f_l)/\rho$ uncorrelated samples of $\mathbf{V}(\omega;\boldsymbol{\theta})$. For $N > \tilde{N}$, the modified CRB of Figs. 8–9 is obtained from (25) with the distinction that $\omega_k = 2\pi k/(\tilde{N}T_s)$. In this case the modified CRB hence corresponds to the GCC error variance (53). For $N < \tilde{N}$, the modified CRB is obtained from the original expression (25). In practical applications the GCC error variance (53) hence appears to be a more relevant performance bound than the CRB.

Note, to present the results in a fair manner (Figs. 8 and 9), any estimate not satisfying $|\hat{\tau} - \tau_0| < 4 \cdot \sqrt{\text{Modified CRB}}$ was excluded from the calculation of the empirical standard deviation.

We notice further that 1) "CC" performs worse than PHAT and AML when the source signal is colored and 2) the theoretical probability of an outlier agrees well with empirical observations also for colored source signals, as long as PHAT or AML is applied and as long as N is large (i.e., $N \gg 2(f_u - f_l)/\rho$).

From Figs. 8 and 9, we also draw the conclusion that AML does not perform better than PHAT. The possible advantage of AML is instead that it offers simultaneous processing of several microphone signals in an effective way.

VI. Conclusion

Several studies in the literature have indicated that the problem of localizing acoustical sources in the presence of room reverberation is difficult, and there has been an interest in analyzing the performance when room reverberation is present.

In the first part of the paper, we applied the theory of statistical room acoustics to develop a statistical model that explains the effects of room reverberation. In the second part we applied the reverberation model to perform a statistical analysis. The Cramér-Rao lower bound for the variance of the estimated time-delay was derived, and a statistical analysis of GCC-based time-delay estimation was given. In this context, our contributions are 1) an explicit expression for the GCC error variance has been derived and 2) the probability of an anomalous estimate has been established.

The theoretical results were evaluated in a number of numerical experiments, where the image method was applied to simulate the room transfer function.

The performed statistical analysis is by no means exact; to be able to perform the derivations several simplifying assumptions were introduced. For our specific numerical examples, the theoretical error variance of GCC showed a good agreement with the empirical results as long as 1) the reverberation time was greater than 0.07 s, 2) $d > 0.2$ m, and 3) the distance between the source and the microphones is larger than 1.5–2 m. For reverberation times shorter than 0.07 s, the room impulse response contains "too few" echoes for the assumption of diffuse sound to be applicable. For small source-microphone distances, the presence of early dominant reflections invalidates the assumption of diffuse sound.

Regarding the derived CRB, it seems reachable only for small sample sizes. This is most likely an effect of an unrealistic assumption on Gaussianity.

Empirical observations together with the presented theory show that GCC-based localization techniques are reliable only for relatively large values of the signal to reverberation ratio (SRR). Especially cumbersome is the fact that GCC-based localization techniques tend to produce large errors (outliers) unless the SRR is large. In our numerical examples, $\mathrm{SRR} > 0$ dB typically guarantees that the probability of outliers is tolerable.

References

[1] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput., Speech, Lang.*, vol. 11, no. 2, pp. 91–126, Apr. 1997.

[2] J. E. Adcock, M. S. Brandstein, and H. F. Silverman, "A closed-form estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 45–50, Jan. 1997.

[3] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proc. ICASSP*, Atlanta, GA, 1996, pp. 921–924.

[4] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, pp. 1223–1225, Aug. 1987.

[5] W. R. Hahn and S. A. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inform. Theory*, vol. 19, pp. 608–614, Sept. 1973.

[6] C. Wang and M. Brandstein, "Multi-source face tracking with audio and visual data," in *Proc. IEEE 3rd Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, 1999, pp. 169–174.

[7] C. Wang and M. S. Brandstein, "A hybrid real-time face tracking system," in *Proc. ICASSP*, Seattle, WA, 1998, pp. 3737–3740.

[8] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 148–152, Mar. 1996.

[9] J. P. Ianniello, "High-resolution multipath time delay estimation for broadband random signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 320–327, Mar. 1988.

[10] T. Gustafsson, B. Rao, and M. Trivedi, "Analysis of time-delay estimation in reverberant environments," in *Proc. ICASSP*, Orlando, FL, 2002.

[11] B. D. Radlović, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 311–319, May 2000.

[12] H. Kuttruff, *Room Acoustics*. New York: Wiley, 1973.

[13] M. R. Schroeder, "Frequency correlation functions of frequency responses in rooms," *J. Acoust. Soc. Amer.*, vol. 34, no. 12, pp. 1819–1823, 1963.

[14] J. B. Allen and A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[15] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986.

[16] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, pp. 320–326, Aug. 1976.

[17] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1783–1795, Oct. 1990.

[18] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[19] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[20] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1997.

[21] E. E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 1321–1324.

[22] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event localization," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 288–292, May 1997.

[23] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 998–1003, Dec. 1982.

Tony Gustafsson (M'01) was born in Värnamo, Sweden, in 1969. He received the M.S. degree in electrical engineering from Chalmers University of Technology, Sweden, in 1994, and in 1999 he received the Ph.D. degree in signal processing from the same university.

From 1999 to 2000, he was a Postdoctoral Researcher at the University of California at San Diego, La Jolla. Presently he is a Research Engineer at Switchcore Corporation, Göteborg, Sweden.

Bhaskar D. Rao (F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, in 1979, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively.

Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor in the Electrical and Computer Engineering Department. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human–computer interactions.

Dr. Rao has been a member of the Statistical Signal and Array Processing Technical Committee. He is currently a member of the Signal Procesing Theory and Methods Technical Committee.

Mohan M. Trivedi (S'76–M'79–SM'86) received the B.Tech. degree in electronics from BITS, India, in 1974, and the M.S. and Ph.D. degrees from Utah State University in 1976 and 1979, respectively.

Since 1995, he has been with University of California at San Diego, La Jolla, where he is currently a Professor in the Electrical and Computer Engineering Department. He and his team are engaged in a broad range of research studies in active perception and novel machine vision systems, and distributed video networks. Noted accomplishments of his team include: active vision systems for depth extraction; multiple camera based human body modeling and movement analysis; a distributed interactive video array for wide area tracking and event-based serving; virtual view synthesis; and 3-D scene modeling using omni-directional video networks.